PCT/DK 2004/000526

DK04/526

Kongeriget Danmark

Patent application No.: PA 2003 01136

Date of filing: 06 August 2003

Applicant:
(Name and address)

Frank Uldall Leonhard
Louisevej 13
DK-2800 Lyngby
Denmark

Title: Auditory perceptual pulse analysis

IPC: G 10 L 101:00

This is to certify that the attached documents are exact copies of the above mentioned patent application as originally filed.

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH
RULE 17.1(a) OR (b)

Patent- og Varemærkestyrelsen
Økonomi- og Erhvervsministeriet

14 July 2004

Susanne Morsing

PATENT- OG VAREMÆRKESTYRELSEN

Auditory Perceptual Pulse Analysis

Abstract

Presented is a new analysis model called Auditory Perceptual Pulse Analysis (APPA) because the model is based on some mixed mathematical and physical phenomena that seem to correspond very much to how the human ear perceive sound. The phenomena are a direct result of the behaviour of transient response of a system. The response reacts in practise as a pulse and the pulse is analysed in auditory frequency channels. The outputs of the channels are fluctuating pulses. The time period of the half-wave fluctuations – ripples - seems very important for the sound picture. It would be very tempting to make a Fourier analysis but the ripples are often composed of more than one frequency, which affects the period time of the ripples. One frequency rides on the back of the other so to speak. Therefore the ripple period is analysed directly. The ear seems to have two frequency ranges where the ripple time is analysed differently: A low range up to 800-1000 Hz where the ripple time is measured from the start of the ripple to the top or to the end; and a high range above 1200-1400 Hz where the ripple time is measured from the top of one ripple to the top of the next ripple. In the low range some vowels are identified and called deep vowels and in the high range high vowels identified. The deep vowels have a representative in the high range that is identified by the fluctuation of the dynamic energy of the output of the auditory channels. Examples of high as well as deep vowels are analysed and it is illustrated that APPA is very noise insensitive identifying vowels.

Introduction

The fundamental basis for auditory perception has been assumed to be short time Fourier transformation. This has been the case because the construction of the cochlea points in that direction. This assumption has however caused several unexplainable phenomena. Among them it has never been revealed how the human ear perceives vowels and in general how it perceives sound "pictures".

The ear is originally developed as a warning system that had to warn against enemies that try sneaking toward you, and it is typically the sound from breaking twigs that gives you a warning. Such a sound is a pulse and it can have a very short duration, which is not expedient in relation to a Fourier transformation because the information contained by the pulse will be averaged through the analysis. Also the frequency spectrum of a signal has the weakness that outstanding frequencies in the spectrum might come from different sources without the possibility to tell. This is a big problem when analysing vowels in speech recognition, where background noise might be interpreted as false formants.

Fourier transformation is a mathematical tool that eliminates the time dimension and is therefore by nature not suitable for pulse analysis. The fact that pulses are very important the dynamic behaviour of the signal is very important and the method for analysis must be based on tools that reflect the physical conduct of the cochlea. Auditory Perceptual Pulse Analysis (APPA) is a time frequency analysis that tries to reflect the approach the human ear.

Pulse Analysis

A pulse is created by an abrupt force hiding a system. It might be a fault on a gear wheel in a gearbox, and each time the fault on the wheel is messed a pulse is generated by the force caused by the fault. Or it could be the pressure explosion created by the vocal cords at a voiced phoneme in speech. That the force is an abrupt force means that the transient response of the system is very important.

The full information hold in the pulse is given through the duration of the pulse. If we have a rectangular pulse we have the following Fourier Transform:

$$f(t) = 1 \qquad 0 \le t \le \tau$$

$$= 0 \qquad \text{Otherwise}$$

$$F(\omega) = \int_{-\infty}^{0} f(t)e^{-j\omega t}\,dt + \int_{0}^{\tau} f(t)e^{-j\omega t}\,dt + \int_{\tau}^{\infty} f(t)e^{-j\omega t}\,dt$$

$$F(\omega) = \int_{0}^{\tau} f(t)e^{-j\omega t}\,dt$$

The Fourier integral can be divided into three terms, where only one term contributes to the Fourier transform, and it is the integral through the duration of the pulse. In other words integration through longer time than the duration of the pulse does not add anything to the spectrum of the pulse. This is in fact a plain reflection, but it is rather important.

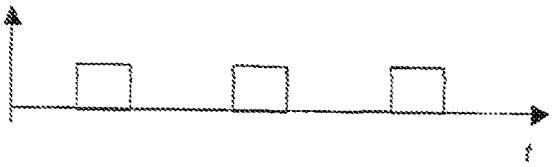


Fig. 1. A pulse train consisting of rectangular pulses.

If we have a pulse train as shown in fig. 1 and the Fourier integral is longer than the duration of the pulses but shorter than the period between the pulses the spectrum will be a continuous spectrum, and the full information of the spectrum of the pulse will be present. If the integration time is longer and contains a sufficient number of pulses the spectrum will be discrete and contain the harmonics of the fundamental frequency. The envelope of the spectrum is equal to the continuous spectrum of a single pulse and the frequency step is equal to the reciprocal time period between the pulses. By integrating through longer time we gain information about the time period between the pulses, but we lose information about the spectrum of the individual pulse because we only have information about the harmonics and not the full spectrum. By pulse analysis the period between the pulses can be determined simply by measuring the time between the pulses.

The physical interpretation of signal that is a pulse train is that the individual pulse is an expression of an occurrence and the period between the pulses is therefore an expression the time between the occurrences. In the case of fault on a gear wheel in a gearbox the spectrum of the individual pulse describes characteristic of the fault and the period between the pulses describes the rotational speed of the wheel.

Auditory Perceptual Pulse Analysis

By a Fourier transform the time domain is eliminated and therefore a Fourier transform is not optimal for describing dynamic conditions in signals. A better method is to divide the auditory frequency interval into a number of auditory channels by means of a filter bank of band-pass filters. If the band-pass filters have broad frequency bands they will have a short impulse response and the time resolution is then high.

It has been assumed for many years that the cochlea of the ear is divided into a number of auditory frequency channels where the signal is analysed [Zwicker 1961]. The methods used were however focused on quasi steady state frequency analysis, among others to be able to trace the formants in speech signals. Another example is Zwicker's Loudness model [Zwicker 1999]. An exception, which is described by F. Leonhard [Leonhard 1993] and [Leonhard 2002], considers abrupt changes in the energy, but a closer relation to how the sound picture is perceived is not described.

Idealized a pulse is an impulse response of a system and contains damped eigenfrequencies, which can be described by poles of the system, and it is in fact a kind of footprint of the system. The objective for a pulse analysis is to get as accurate and much information out of the pulse and it seems to be the case for auditory perception. In the following the focus will be put on analysing the "colour" of the sound picture or vowels, which can be considered as special "colours" of the sound.

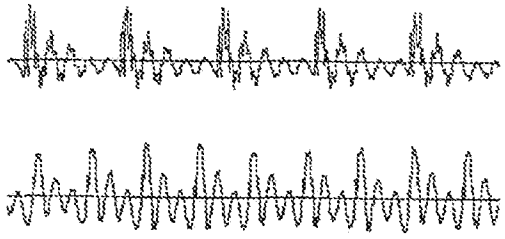


Fig. 2. 50 msec of the vowel "æ:" as in "had" pronounced by a mal and a female.

Fig. 2 shows 50 msec of the vowel "æ:" as in "had" pronounced by a male (the upper signal) and a female (the lower signal). As it is seen the signals contains damped frequencies – most obvious for the male. It is also clear that the interval between the pulses is about twice as long for the male compared to the female.

To have a high time resolution the band-pass filters in the auditory channels have to have broad frequency bands. In fig. 3 10 msec of vowel "æ:" spoken by a male (about one pulse) and the output of 6 band-pass filters shown. The pass-bands for the 6 filters are: 150-450, 300-900, 450-1350, 1400-2800, 2000-4000, and 2800-5600. The gain of the filters is adjusted relatively to compensate for the frequency dependency sensitivity of the ear, and is set to 1, 1, 1, 3, 3, and 2.

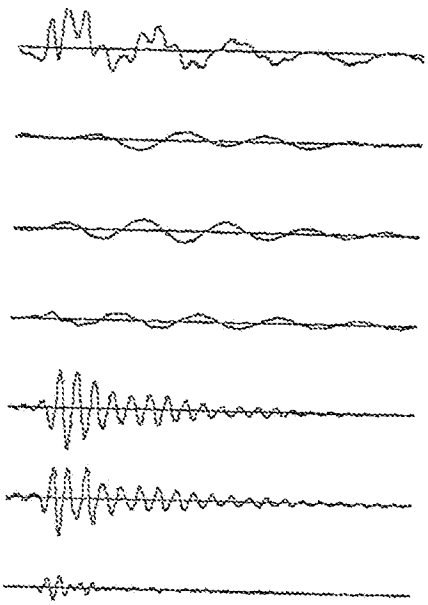


Fig. 3. 10 msec (one pulse) of the vowel "æ:" (male) and output from the 6 band-pass filters.

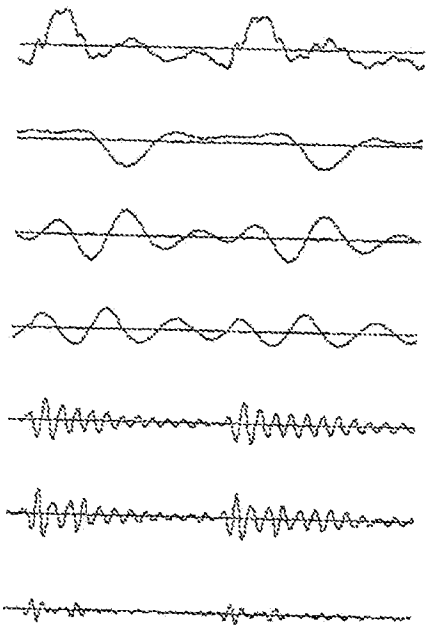


Fig. 4. 10 msec (two pulses) of the vowel "æ:" (female) and the output from the 6 band-pass filters.

Fig. 4 shows 10 msec (about two pulses) of the vowel "æ:" pronounced by a female, and the output from the 6 band-pass filters. Compared to the vowel spoken by the male the period between the pulses is only the half – about 5 msec for the female and 10 msec for the male. However the periods of the fluctuation of the 4th and 5th band-pass filter are very similar. In other words the fluctuation is independent of the pitch period.

On that background there are two phenomena that are of interest in the pulses. One is the nature of the fluctuation. The time period of the half-wave fluctuations – ripples - seems to be very important for the sound picture. The other is the progress of the instantaneous energy through the duration of the pulse. It would be very tempting to make a frequency analysis by means of Fourier transform to analyse the ripples, but the ripples are often composed of more than one frequency, which affects the period time of the ripples. One frequency rides on the back of the other so to speak. Therefore a better method would be to measure the time period of the ripple in a kind of ripple analysis.

Ripple Analysis

As it appears from fig. 3 the output from the filters is rather periodic. This could lead to the assumption that fluctuation only consists of one frequency. But band-pass filters are very broad banded and the pulse is a transient response. This means that the fluctuation often is formed by more than one frequency. In vowels it is typically the second formant that dominates but it is affected by one of the other formants.

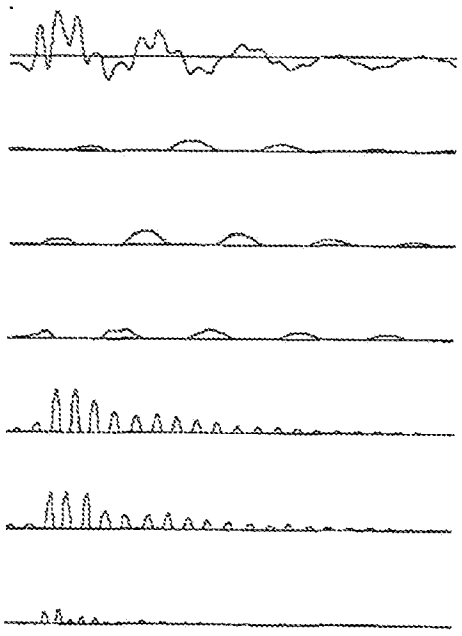


Fig. 5. One pulse from a speech signal and output from 6 band-pass filters that are recilfied, which is called the ripple signal.

To simplify the analysis the output of the band-pass filters are half-wave rectified, and these signals are called ripple signals.

There is every indication that the ear is divided two areas were the ripple is analysed different. In lower frequency range up to 1,000-1,500 Hz, it looks like the ripple is analysed by measuring the time from crossover to the top of the ripple or to the next crossover where the signal goes zero. Vowels detected by this method will be called deep vowels. In the frequency range above 1,000-1,500 Hz the ripple is measured from the top of a ripple to the next. These vowels will be called high vowels. This is anyway convenient in digital signal processing because it increase the resolution about 4 times in the upper frequency range. It is easier to measure the ripple if the output from the band-pass filters is half-wave rectifies. Fig. 5 shows the rectified signals of the vowel "æ:". The corresponding signal in the ear will properly be slightly low-pass filtered, but it is not done here.

The "colour" or vowel is then analysed by measuring the ripple for each auditory channel in a sufficient time frame and sort them in time bins each representing a predefined duration.

To validate the method 4 vowels are analysed spoken by a female and a male. The vowels are "i:" as in "heed", "æ:" as in "had", "a:" as in "hod", and "u:" as in "who'd".

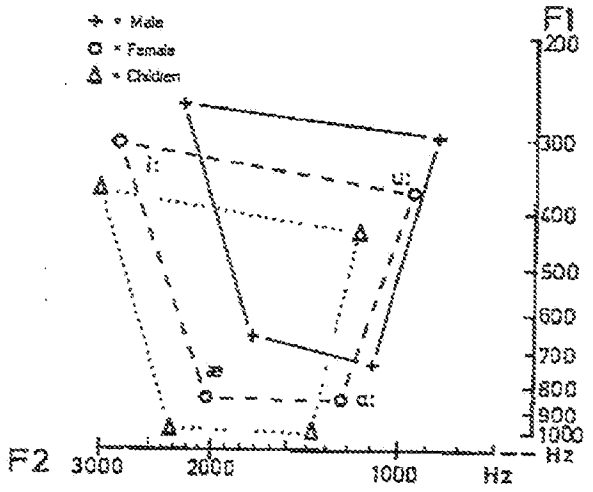


Fig. 6. "i:", "æ:", "a:", "u:", as in "heed", "had", "hod", "who'd".

In phonetics these vowels are regarded as cornerstones in the map of vowels. On fig. 6 [Thorsen 1978] shows the average placement of the first formant F1 and the second formant F2 of the 4 vowels pronounced by males, females, and children. As it is seen there is a fairly big displacement between males, females, and children. This displacement is another problem in speech recognition.

In this analysis the above-mentioned 6 band-pass filters define the auditory channels. The first 3 will detect deep vowels and last 3 will detect high vowels. The deep vowels are detected as the rise time from zero to the top of the ripple, which is about a quarter of a period, while the high vowels are detected as the period from one top

to the next. To get a homogeneous picture of the time bins for the deep and high vowels, the rise time of the deep vowels is multiplied by 4 and called the period of the beep vowel.

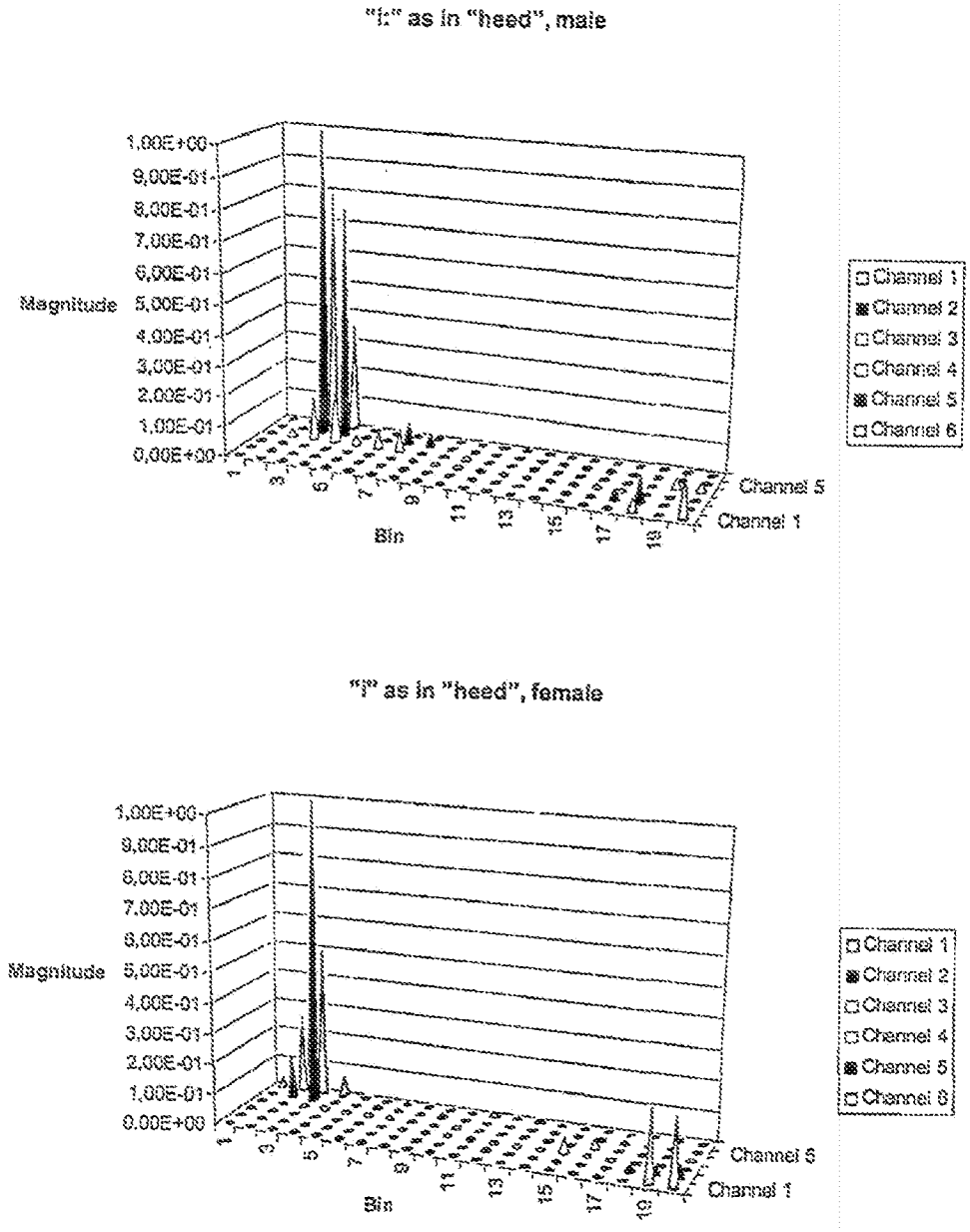


Fig. 7. i: as in heed, pronounced by a male and a female.

The minimum and maximum bound of the bank of bins are in this analysis chosen to 0.2 and 5 msec. The number of bins is selected to 20 and the width the time bins is logarithmic. The time frame of the analysis is selected to 30 msec. In the frame all

ripple periods are measured and sorted according to the channel and time bin. For each channel and time bin all the magnitudes of the ripples that are detected are accumulated. The accumulated magnitudes are normalized to the maximum accumulated magnitude for each analysis and just called the magnitude.

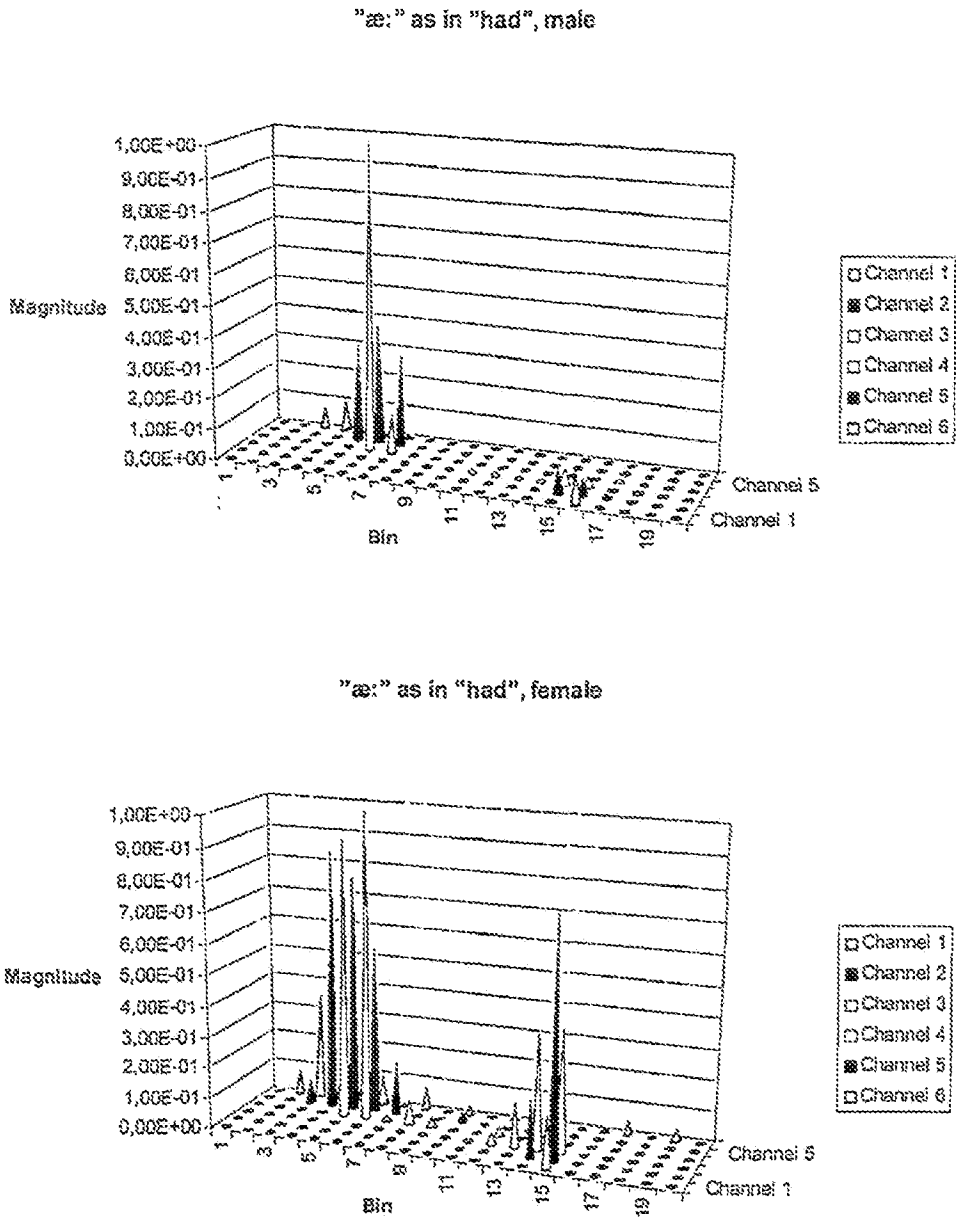


Fig. 8. æ: as in had, pronounced by a male and a female.

Fig. 7 shows the result of the analysis of "i:" as in "heed" pronounced by a male and a female. In both cases the maximum magnitude is found in channel 5, bin 3. That

means that the maximum magnitude is found in the frequency band 2000-4000 Hz and with a ripple period between 0.276 and 0.381 msec.

Fig. 8 shows the result of the analysis of "æ:" as in "had" pronounced by a male and a female. In both cases the maximum magnitude is found in channel 4, bin 6. That means that the maximum magnitude is found in the frequency band 1400-2800 Hz and with a ripple period between 0.447 and 0.525 msec.

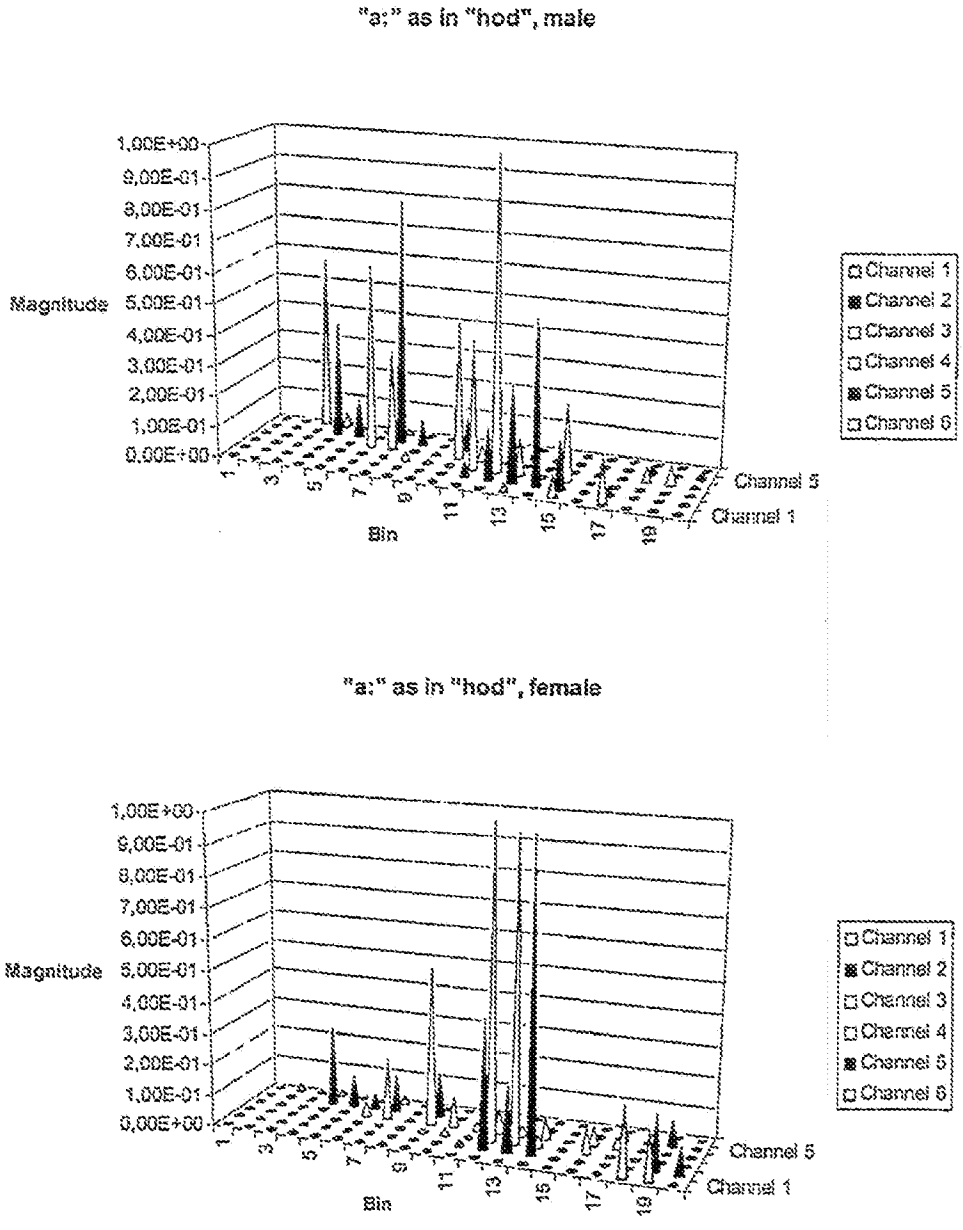


Fig. 9. a: as in hod pronounced by a male and a female.

Fig. 9 shows the result of the analysis of "a:" as in "hod" pronounced by a male and a female. In both cases the maximum magnitude is found in channel 3, bin 12. That means that the maximum magnitude is found in the frequency band 450-1350 Hz and with a ripple period between 1.175 and 1.380 msec.

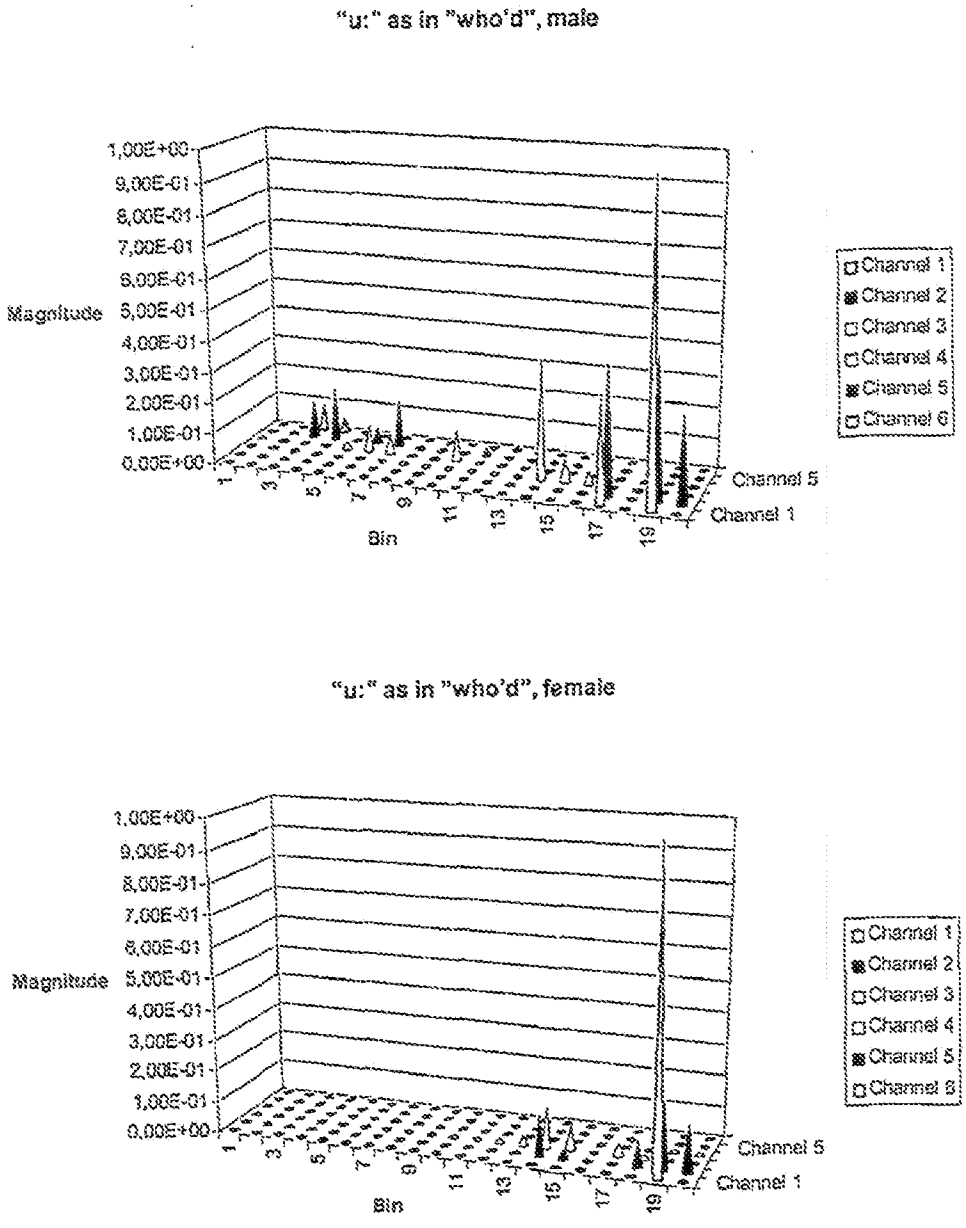


Fig. 10. u: as in who'd pronounced by a male and a female.

Fig. 10 shows the result of the analysis of "u:" as in "who'd" pronounced by a male and a female. In both cases the maximum magnitude is found in channel 1, bin 19. That means that the maximum magnitude is found in the frequency band 150-450 Hz and with a ripple period between 3.624 and 4.257 msec.

The ripple analysis of the 4 vowels spoken by a male and a female, shows clearly that ripple period reflect the vowel independent of the pitch. The pitch of voice of the female was about twice the pitch of the voce of male without any influence on the ripple analysis.

The ripple analysis of vowels is extremely noise insensitive. Fig 11 shows the ripple analysis of "u:" pronounced by the female as shown on fig. 10 but added –5 dB noise from a car. Compared to fig. 10 there is only very little difference.

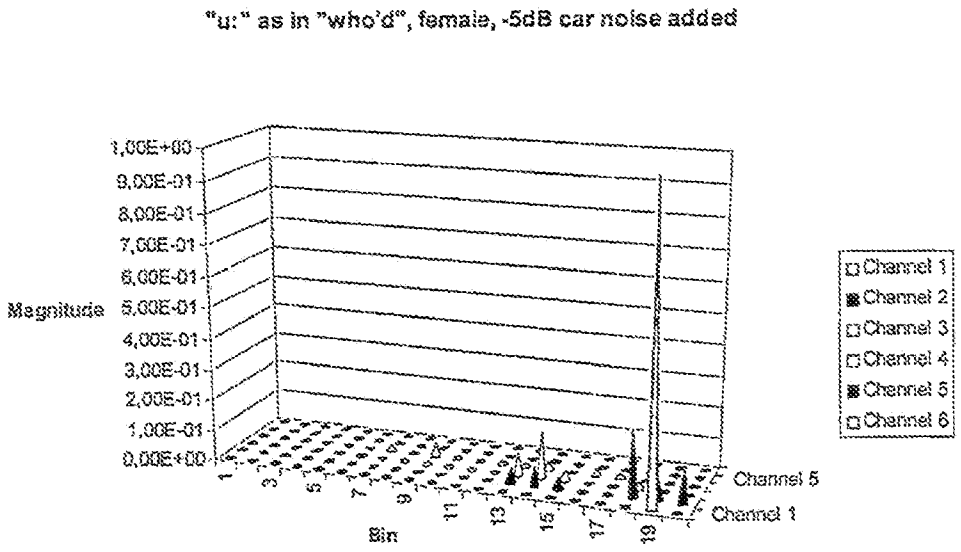


Fig. 11. u: as in who'd pronounced by a female added – 5 dB care noise.

In appendix A the ripple periods of some Danish vowels (high vowels) are listed. In the above analysis the resolution of time bins is 17.5%, but as it is seen from the appendix it has at least has to better than 10% and properly as good 3-5%.

Dynamic Energy Analysis

Dynamic energy analysis (some times also called envelope or transient analysis) is an analysis where the dynamic instantaneous energy is analysed in the auditory channels. The method might be based on a half- or full-wave rectification followed by a low-pass filtration. In general the more abrupt the pulse is; the more "sharp" the sound picture will be. Fig. 12 shows the dynamic energy of the vowel "æ:" pronounced by a male. By nature the most abrupt changes will be present in the high frequency range.

The some of the deep vowels have the problem that under some conditions they are not detectable in the low frequency range below about 1000 Hz. It might be if people are talking through a phone or under special noise circumstances. Under these circumstances these deep vowels are represented in the high frequency range as dynamic energy fluctuations.

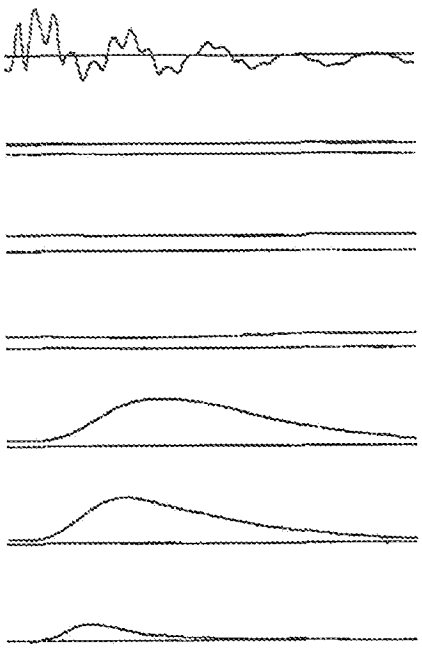


Fig. 12. Dynamic energy of one pulse of the vowel "æ:"

This modulated presentation seems to be best represented by the derivative of the dynamic energy. Fig. 13 shows the derivate of the dynamic energy of the vowel "u:" high-pass filtered with a cut-off frequency at 1400 Hz and fig. 14 shows the derivate of the dynamic energy of the vowel "a:" also high-passed filtered at 1400 Hz.

What criteria that have to be met to decide when a ripple or a dynamic (derivate of the dynamic energy) analysis is valid in the high frequency range have not been studied in details yet. In the case of dynamic analysis one think is however clearly; the period of the dynamic fluctuation has to be in a given interval. Other factors are that in some cases the ripple periods have a certain scatter and therefore not perceived as high vowel, and in other cases the ripple period is longer than the periods valid for high vowels.

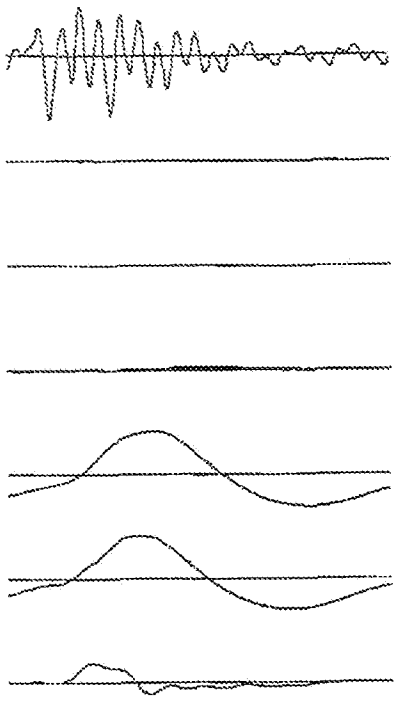


Fig. 13. The derivate of the dynamic energy of one pulse of the vowel "u:" high-pass filtered at 1400 Hz.

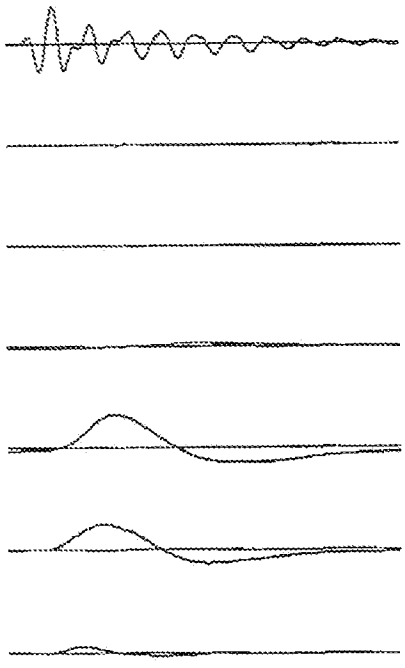
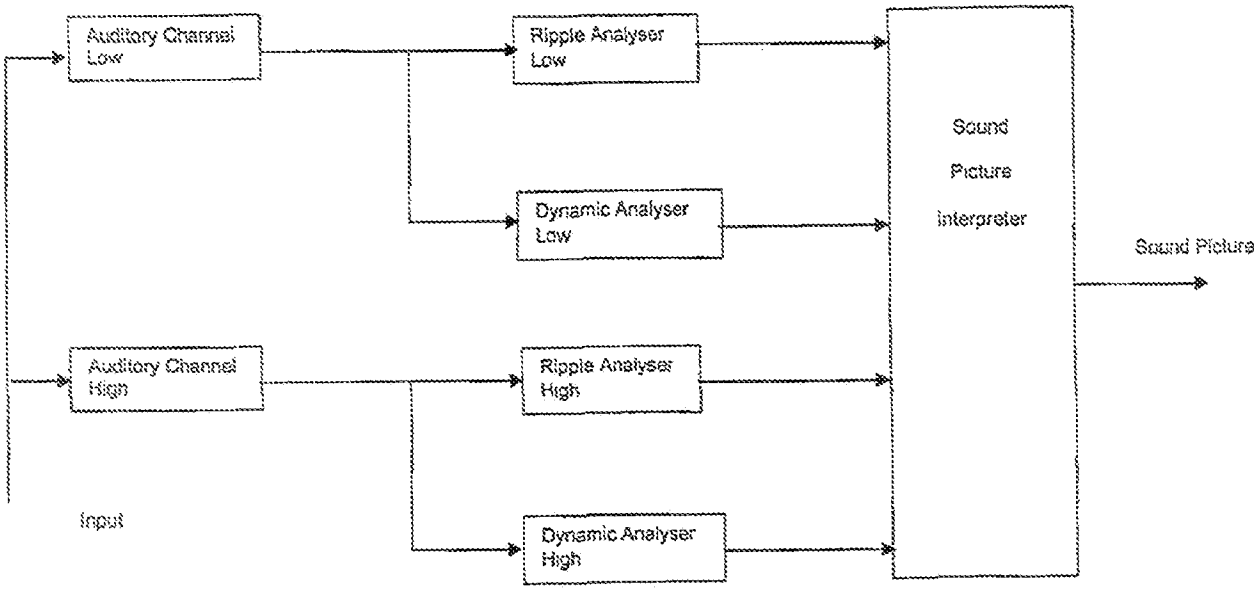


Fig. 14. The derivate of the dynamic energy of one pulse of the vowel "a:" high-pass filtered at 1400 Hz.

APPA Model



The model is based on the assumption that the cochlea is very broad-banded when it is relaxed and acts like adaptive filters if it is actuated by sinusoids and gets more narrow-banded around the frequencies.

Auditory Channel Low is a filter bank containing band-pass filters in the range from 150 to 1500 Hz. It might only contain one filter with pass-band from about 200 to 1000 Hz.

Auditory Channel High is a filter bank containing band-pass filters in the range from 1000 to 6000 Hz. It might only contain one filter with pass-band from 1500 to 4000 Hz.

The full model will have an Auditory Channel Ultra High, which is a filter bank containing band-pass filters in the range 4000 to 20000 Hz.

The Ripple Analyser Low analyses the ripple period by measuring the time from a positive ripple starts and it has its maximum or it goes to zero again. The ripples are sorted after the length of their periods in bins and their magnitudes are accumulated in the respective bin through a specified frame.

The Ripple Analyser Low analyses the ripple period by measuring the time from a positive ripple has its maximum to the next ripple has its maximum, or it could be measured from a ripple starts to the next starts. The ripples are sorted after the length of their periods in bins and their magnitudes are accumulated in the respective bin through a specified frame.

The Dynamic Analyser Low analyses the dynamics by detecting the derivate of the dynamic energy. The magnitude and duration of positive going fluctuations of the derivate are measured and are sorted after the length of their periods in bins and their magnitudes are accumulated in the respective bin through a specified frame.

The Dynamic Analyser High analyses the dynamics by detecting the derivate of the dynamic energy. The magnitude and duration of positive going fluctuations of the derivate are measured and are sorted after the length of their periods in bins and their magnitudes are accumulated in the respective bin through a specified frame. Also the period between the pulses (the pitch) throughout the frame might be analysed.

The result of the analysis might be presented as graphs but it might also be interpreted into more specific sound pictures.

Discussion

It is shown that the ripple in auditory channels has unique characteristics for four different vowels spoken by a male and female and characteristics of the ripple seems to be very important for how the ear perceives the colour of the sound or identifying vowels. The ripple is coursed by transient response of systems, which means that it is phenomenon that consists of a special time frequency relation. It might be assumed that the auditory channels of the cochlea are very broad-banded when it is relaxed or affected by broad-banded transient signals. If the cochlea is affected by a steady state signal consisting of one ore more sinusoids it might act like adaptive filter and be more narrow-banded around the frequencies.

The band-pass filters used in this analysis are the first estimate. Standard Butterworth filters were chosen to give a rapid illustration of the ability of the technique, and there are room for improvement. A big difference between a transient response and a steady state response is that in a steady state response only frequencies of the forced signal are present. A transient response consists as well of the frequencies of the forced signal as of the Eigenfrequencies of the system. That means that the poles of the filters that form the auditory channels will affect the ripple of the pulses but it shall be as little as possible. In this analysis Butterworth filters have been use but the best choice would most probably be filters with real poles and with an impulse response that are critical damped as suggested by Leonhard [Leonhard 2002]. Butterworth filters have poles in the complex plane and some of them are fairly close to the imaginary axis, and they will interfere with the signal and in same cases give false ripples.

In this analysis the bounds for the low frequency range have been chosen to 150 Hz for the lower and 1350 Hz the upper bound, and for the high frequency range to 1400 Hz for the lower and 5600 Hz for the upper bound. Studies of a lot of different vowels indicates that a choice might be 200-300 Hz for the lower bound and 1300-1400 Hz for the upper bound in the low range, and 1300-1400 Hz for the lower bound and 3500-4000 Hz for the upper bound in the high range. Further it seems that if the period between two ripples is longer than about 0.7 msec the ripple time from the start to the top of the ripple that is important. Is the period shorter than 0.7 msec then it is the period that is important. How many channels, the bandwidth of the filters and their placement has to be further investigated by evaluating tests with different set of filters and their affect on the ripple. Especially in the interval between the low and high range the selection of filters are critical.

The time bounds in this analysis have been chosen to the total interval from 0.2 to 5.0 msec divided into 20 bins. The interval between the bounds for each bin of the 20 bins is calculated to be logarithmic. To make it easier to compare the ripple time between the deep and high vowels the ripple time of the deep vowels is multiplied by 4. The time where the ripple starts to the top can be viewed as a quarter of a period and then it is comparable to the ripple period of the high vowels.

The time resolution of different bins is 17.5% with the chosen figures as mentioned above. As it is seen of appendix A showing some Danish high vowels this resolution is too coarse for the high vowels. It has at leased to be better than 10% and properly around 2-3%. On the other hand the total interval can be divided into subsections. The total interval for the high vowels might only be about from 0.25 to 0.8 msec. The total interval of interest for the deep vowels from the start to the top is around from 0.2 to 1.1 msec, but it does not look like that the requirement to the time resolution is as high as for the high vowels.

It is not possible to give more details of the figures for the representation of the dynamic energy fluctuation of deep vowels before a higher time resolution has been implemented. It will also be preferable that critical damped filters with real poles are implemented for the auditory channels.

Applications

APPA is of cause very applicable for speech analysis. Even vowels that are whispered can be identified. It will also be applicable for establishing methods for analysing the intelligibility of speech, as well as methods for analysing sharpness and roughness of a sound picture, and the pitch of speech or music. APPA will also be able to lead to speaker independent and much more robust speech recognition and more efficient speech compression with higher quality in narrowband telecommunication than known today.

Other applications areas are analysis of all kind of mechanical faults in moving machinery. It might be defects of ball bearings or of gear wheels in a gearbox. Also supervising mechanical establish or in general processes where the sound picture reveals quality or faults. It might be cutting in wood or metal, or it might be defining a specific sound picture of a product as the sound of a motorbike or the sound of a car door slamming.

APPA will also be applicable for measuring the quality of audio devices of all kind and set up standards for the quality.

In the medical world there are cases the shape of never pulses have interest and APPA will be able to add very important information to the diagnoses of illness of nerve systems. Also analysis of heart beets and sounds from lungs will bring important information for medical diagnosis.

References

[Leonhard 1993] Frank U. Leonhard, "Method and System for Detecting and Generating Transient Conditions in Auditory Signals", EP 0737351, April 1993.

[Zwicker 1961] E. Zwicker, "Subdivision of the audible frequency range into critical bands", Journal of the Acoustical Society of America, 33, page 248-249, 1961.

[Zwicker 1999] E. Zwicker, H. Fastl, "Psychoacoustics", Second Updated Edition, Springer, 1999.

[Seneff 1988] Stephanie Seneff, "A joint synchrony/mean-rate model of auditory speech processing", Journal of Phonetics (1988) 16, 55-76.

[Thorsen 1978] Nina Thorsen & Oluf Thorsen, "Fonetik for sprogstuderende", Institut for Fonetik, Københavns Universitet, 3. reviderede udgave, 1978.

[Leonhard 2002] Frank Uldall Leonhard, "Quality Control of Electro-acoustic Transducers", WO 02/25997, Marts 2002.

Appendix A

Approximated ripple periods for some Danish high vowels:

"i", hive	0.30 msec.
"e", hele	0.34 msec.
"æ", hæle	0.38 msec.
"y", hyle	0.43 msec.
"a", hale	0.48 msec.
"ø", høne	0.55 msec.